7

Correlation and concordance

7.1 CORRELATION IN BIVARIATE DATA

Nonparametric correlation is concerned largely with paired observations consisting of ranks. The ranks may be the primary data or they may be derived from continuous measurements. In a parametric context with measurement data, the most widely used indicator of correlation is the **Pearson product moment correlation coefficient.** For n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ this coefficient is

$$r = c_{xy} / [\sqrt{(c_{xx} c_{yy})}] \tag{7.1}$$

where $c_{xy} = \Sigma(x_i y_i) - (\Sigma x_i)(\Sigma y_i)/n$, $c_{xx} = \Sigma(x_i^2) - (\Sigma x_i)^2/n$, $c_{yy} = \Sigma(y_i^2) - (\Sigma y_i)^2/n$, all summations being over the subscript i. The Pearson coefficient is particularly relevant to samples from a bivariate normal distribution where it is an appropriate estimate of the population correlation coefficient ρ, but it is used in practice in a wider context as a measure of linear association in the sense that r takes the value $+1$ or -1 if there is a straight line relationship between x and y. Generally values of r close to $+1$ or -1 imply a near-linear relationship between continuous x and y. If both x and y have a normal distribution, values near zero imply independence. In general, if x and y are independent r will be close to zero, but the converse is not true; there may well be some nonlinear relationship between x and y.

Desirable properties of any correlation coefficient are that its values should be confined to the interval $(-1, 1)$ and that lack of association implies a value near zero. Values near $+1$ imply a strong positive association (i.e. high values of y are associated with high values of x and low values of y with low values of x) and values near -1 imply a strong negative association (i.e. high values of y are associated with low values of x and low values of y are associated with high values of x). For the Pearson coefficient, $r = \pm 1$ implies linearity, but for rank coefficients values of ± 1 need not imply linearity in continuous data from which these ranks may have been derived. They do imply a property called **monotonicity**. This means that as x increases y increases (monotonic increasing) or as x increases y decreases (monotonic decreasing). For rank correlation the value 1 implies strictly increasing monotonicity, the value -1 strictly decreasing monotonicity. Values near ± 1 imply a situation approaching monotonicity.